Unsupervised IDS: Exploring the Measures

P. S R Mty
Asst.Professor, dept. of IT
Vishnu Institute of Technology
Bhimavaram, Andhra Pradesh, India

R. K Kumari
Assoc.Professor, dept. of CSE
Krishna University
Machilipatnam, Andhra Pradesh,

N. Shailja
Professor, dept. of ECE
JNTUK
Kakinada, Andhra Pradesh, India

***Abstract*—This paper investigates various Similarity/Dissimilarity measures for Intrusion Detection Problem. In this paper we implemented an offline Anomaly based IDS using agglomerative and partition based clustering algorithms with selected Similarity/Dissimilarity measures. In unsupervised learning labeling the clusters is an important task. This paper employed two cluster labeling algorithms, SNC labeling algorithm and "labeling clusters using class representative objects". This work is evaluated using KDDCup 99 dataset.**

Keywords—distance/similarity measures; hierarchical clustering; Anomaly Intrusion Detection System, K-Medoids, Labeling Clusters;

I. INTRODUCTION

A network intrusion or a security breach is an activity which causes unwanted access of network resource or unauthorized modifications to information. Now a day's Intrusion Detection Systems have become an inevitable second line of defense which comes into play when the traditional security mechanisms fail and an intrusion happens in the network. Network Intrusion Detection Systems are treated as two categories, Signature based and Anomaly based. Signature based IDS compares the network traffic with the attack signatures to detect intrusions. Anomaly based IDS establish a base line (normal behavior of the network) and any deviations from the base line are treated as attacks. In this paper we implemented an offline Anomaly based IDS and presented the experimental results. Unsupervised Anomaly detection algorithms make the assumptions that, the number of anomalies is much less than the number of normal instances and the anomalies are statistically different from the legitimate connections [7].

A cluster contains a group of objects which are similar to each other, and dissimilar when compared to objects in another cluster. Most of the clustering algorithms depend on the notion of similarity, for which several similarity/dissimilarity measures are available. Similarity/Dissimilarity measures which we described in this paper are Euclidean distance, Squared Euclidean, Manhattan distance, Minkowski Distance, Bray-Curtis distance, Ruzicka Similarity index, and Robers Similarity index. We use these similarity measures to cluster the network connections into two groups' normal and attack. The clustering algorithms studied are Hierarchical Clustering and K-Medoids Clustering. We used Similarity Normal Cluster Labeling algorithm proposed by Othman et.al [1], and a semi-supervised approach "labeling clusters using class representative objects", which is based on cluster similarity assumption for labeling the clusters.

In Section II, various similarity/distance measures we selected for our experiments are described. In Section III, the clustering algorithms are described. In Section IV, details of the datasets employed in this paper are discussed. The experimental setup and results are described in Section V. Section VI is about the conclusion and future work.

II. SIMILARITY/DISTANCE MEASURES

Similarity/Dissimilarity measures are fundamental in clustering, pattern recognition, and Information Retrival fields. Lot of research has been going on for identifying new similarity/dissimilarity measures for centruies. Besides the well established Euclidean Distance there exists several similarity/dissimilarity measures which play a prominent role in finding clusters or patterns among the observations. Based on different domains the researchers have proposed several similarity/dissimilarity measures which are specific to their domain or general domain.

A. Theory of Distance Measures

Mathematically a distance is the degree of dissimilarity between two objects. These distances are called as Distance Metrics or Distance measures o sometimes divergences based on weather they satisy the metric properties or not. The required properties for a distance metric are[8],

- *Positivity:* the distance between any two objects is always positive, i.e., dist(x,y) >= 0.
- *Identity of indiscernibles:* The distance between two objects x and y is zero only if x and y are identical i.e., dist(x,y)=0 => x= y
- *Symmetry:* the distance between two objects x and y is always equal to the distance between y and x, i.e., dist(x,y)=dist(y,x).
- *Triangle Inequality:* the distance between x and y should be always less then or equal to the sum of the distance between x and z, and y and z, i.e., dist(x,y)<= dist(x,z)+dist(z,y).

B. Distances

$$\boldsymbol{D} = \begin{bmatrix} (a_{(1,1)}, a_{(1,2)}, a_{(1,3)}, \ldots\ldots, a_{(1,m-1)}, a_{(1,m)}) \\ (a_{(2,1)}, a_{(2,2)}, a_{(2,3)}, \ldots\ldots, a_{(2,m-1)}, a_{(2,m)}) \\ (\ldots\ldots) \\ (\ldots\ldots) \\ (\ldots\ldots) \\ (\ldots\ldots) \\ (\ldots\ldots) \\ (a_{(n,1)}, a_{(n,2)}, a_{(n,3)}, \ldots\ldots, a_{(n,m-1)}, a_{(n,m)}) \end{bmatrix}$$

D represents the data matrix, i.e., Intrusion Detection dataset, where 'n' is the number of connections and 'm' is the number of attributes or dimensions.

Let $p_1=(a_{11},a_{12},a_{13},\ldots a_{1m})$ and $p_2=(a_{21},a_{22},a_{23},\ldots..,a_{2m})$ be the two connections in the intrusion detection data. There are several distance metrics/measures for information consult [3].

1) Euclidean Distance: It measures the geometric distance between two points, based on the fact that the minimum distance between any two points is the length of the straight line joining those two points. Euclidean distance is also known as L^2 distance. Then the Euclidean distance is calculated by using the following expression,

$$d^{e}(p_1, p_2) = \sqrt{\sum_{i=1}^{m}(a_{1i} - a_{2i})^2} \quad (1)$$

2) Squared Euclidean Distance: It is also used as a distance measure as it gives more weight for objects that are far away.

$$d^{se}(p_1, p_2) = \sum_{i=1}^{m}(a_{1i} - a_{2i})^2 \quad (2)$$

3) Manhattan Distance: It is the simple sum of differences of horizontal and vertical components which are measured along axes at right angles between two points' p_1 and p_2. This is also known as Taxicab distance.

$$d^{man}(p_1, p_2) = \sum_{i=1}^{m}|a_{1i} - a_{2i}| \quad (3)$$

4) Minkowski Distance: It is the generalization for Euclidean and Manhattan distances. The formula for the Minkowki distance of order p is given in equation (4). Minkowski Distance is generally called as L_p Norm.

$$d^{mink}(p_1, p_2) = (\sum_{i=1}^{m}|a_{1i} - a_{2i}|^p)^{1/p} \quad (4)$$

5) Bray-Curtis: distance is a modified Manhattan distance i.e., the simple sum of differences measured along the axes which is divided by the sum of features of the objects.

$$d^{bcd}(p_1, p_2) = \sum_{i=1}^{m}|a_{1i} - a_{2i}| / \sum_{i=1}^{m}(a_{1i} + a_{2i}) \quad (5)$$

6) Ruzicka: similarity index is another measure we used in this work

$$S^{ruz}(p_1, p_2) = \sum_{i=1}^{m}\min(a_{1i}, a_{2i}) / \sum_{i=1}^{m} max(a_{1i}, a_{2i}) \quad (6)$$

7) Roberst Similarity Index is given in equation "(7),"

$$S^{rob}(p_1, p_2) = \sum_{i=1}^{m}\left[(a_{1i} + a_{2i})\frac{min(a_{1i}, a_{2i})}{max(a_{1i}, a_{2i})}\right] / \sum_{i=1}^{m}(a_{1i} + a_{2i}) \quad (7)$$

III. CLUSTERING

Clustering by itself is a potential research area, where large number of objects are organized into groups based on the similarity among them. There exists many different ways of defining similarity among the objects, that is why there is no universally accepted unique best way defined for clustering.

Clustering is an unsupervised learning as the class labels are not supplied along with the dataset (observations) for clustering algorithm. Clustering has its applications in almost all fields like biological data analysis, ecological data analysis, multimedia data analysis, social network analysis, finance, engineering and so on.

A. Hierarchical Clustering

Hierarchical Clustering algorithms have two categories, one method starts from singleton clusters and iteratively merge their nearest neighbous to form higher level clusters (agglomerative clustering, bottom-up approach) in another category clustering starts from a big cluster i.e., whole dataset is treated as one cluster and it is partitioned until singleton clusters are evolved (divisive clustering, top-down approach). We used agglomerative clustering method for our experiments.

1) Agglomerative Clustering Algorithm

a) Start: Every observation is a singleton cluster, i.e., $c_i=\{a_i\}$.

b) Find nearest pair of clusters: $\left(min\left(D(c_i, c_j)\right)\right)$

c) Merge the nearest clusters (c_i, c_j) *to form a new cluster* c_{i+j}.

d) Add the merged new cluster c_{i+j} *in the cluster collection C and remove* (c_i, c_j).

Repeat {b,c,d} until one cluster is left. This clustering algorithm produces hierarchical tree of clusters.

B. K-Medoids

K-Medoids is a k-partioning clustering algorithm. K-Medoids is robust, it is not sensitive to outliers. Instead of taking mean value of objects as the centroid, the most centrally located object is selected and that object is called as medoid.

K-Medoids can be applied by using the program PAM (Partitioning Around Medoids)[2]. We choose K-Medoids as it can do clustering by taking the distance matrix as input.

1) K-Medoids Algorithm:

a) Initially select K-objects as the initial representative cluster medoids.

b) Assign each point to the cluster with the closest medoid.

c) Randomly select a non-representative object O_i

d) Compute the total cost 's' of swapping the medoid 'm' with O_i

e) If s<0, then swap 'm' with O_i to form the new set of medoids

Repeat {b,c,d,e} until convergence criteria is achieved.

IV. DATASETS

KDDCup 99 dataset [6] is treated as a standard benchmark dataset which is publicly available for Network Intrusion Detection research. Even there are many critisims on this dataset [4] still it is being used by many researchers for their studies.

KDDCup 99 dataset is prepared by Stolfo et al. [5] from the DARPA'98 data. In this study 10% KDDCup 99 training dataset is used for experiments. All the samples in our experiments are derived from the 10% KDDCup 99 Dataset. The 10% KDDCup 99 dataset contains 494,021 observations with 41 attributes and provided with a class label for every observation.

For experiments we used 2 datasets, for first dataset we selected all the connections whose service type is "smtp", from the 10%_KDDCup 99 training set and we named it as Smtp_data. This dataset contains 9723 connections, in which 9598 connections are 'normal' and 125 connections are 'attacks'.

The second dataset is formed by selecting all the connections whose service type is "ftp_data", from the 10%_KDDCup 99 training set and named it as ftp_dataset.

V. EXPERIMENTS

In this paper we implemented an offline Anomaly based IDS and presented the experimental results. The two clustering techniques Hierarchical Clustering and PAM are implemented and evaluated on datasets. The resulting clusters are labeled using the SNC labeling algorithm described in [1]. On each dataset, each clustering algorithm is applied number of times each time employing a different distance measure. Each clustering algorithm is implemented many times even with the same distance measure by increasing the number of clusters.

The performance of the clustering algorithms based on the distance measures and the labeling algorithms is evaluated using the Detection Rate (DR) and the False Alaram Rate (FAR).

TABLE I. DR AND FAR ON SMTP_DATA USING SNC LABELING ALGORITHM

Algorithms				
	Distance Measures	***FAR(%)***	***DR(%)***	***#Clusters***
Hierarchical Clustering Algorithm (hclust)	Euclidean Distance	2.68	99.2	12
	Manhattan Distance	0.9	99.2	9
	Minkowski Distance (p=0.75)	0.9	99.2	9
	Minkowski Distance (p=1.5)	1.09	99.2	11
	Bray/Curtis Distance	0.59	99.2	5
	Ruzicka Distance	0.59	99.2	8
	Roberts Distance	0.59	99.2	7
K – Medoids (PAM)	Euclidean Distance	1.72	99.2	15
	Manhattan Distance	1.60	99.2	15
	Minkowski Distance (p=0.75)	1.78	99.2	15
	Minkowski Distance (p=1.5)	1.57	99.2	15
	Bray/Curtis Distance	0.20	96.8	10
	Ruzicka Distance	0.20	96	10
	Roberts Distance	0.15	92.8	10

[a.] One attack (satan) was always in a big cluster

We increased the number of clusters until all the attacks are clearly clustred, so in almost all the experiments with smtp_data dataset we achieved near 100% detection rate, i.e., able to label the 124 connections as "attack" out of 125 attack connections, but a few normal connections are labeled as "attack" which raises false alarms. We calculated the rate of the normal connections labeled as "attack" and called it as False Alarm Rate (FAR) which we were able to get a minimum of 0.59 while using Bray/Curtis or Ruzicka or Roberts distance measure with hierarchical clustering at the same time achieving 99.2% detection rate. The results of our experiments are shown in Table I.

TABLE II. COMPARISON OF CLUSTERS WITH GROUND TRUTH (HCLUST USING BRAY/CURTIS DISTANCE MEASURE)

Clusters	**Ground Truth (attacks)**				
	ipsweep	***neptune***	***normal***	***portsweep***	***satan***
1	0	0	2253	0	1
2	0	0	3678	0	0
3	0	0	3349	0	0
4	0	0	261	0	0
5	1	120	57	2	1

[b.] Before applying SNC Labeling algorithm

Table II represents the comparison of the clusters formed using Hierarchical clustering with Bray/Curtis distance measure. In the table we can see that there are five clusters, and by using SNC Labeling algorithm all the connections in cluster 5 are labeled as "attack". After applying the SNC Labeling algorithm the resultant classes are shown in a confusion matrix in table III

TABLE III. CONFUSION MATRIX AFTER APPLYING SNC LABELING ALGORITHM

	Normal	***Attack***	
Predicted Normal	9541	1	9542

	Normal	*Attack*	
Predicted Attack	57	124	181
	9598	125	

[c.] Smtp_data using hclust and bray/Curtis distance method

A. *Experiments with ftp_dataset*

The second data set we used for our experiments is ftp_dataset, we selected all the connections whose service type is ftp_data and named it as ftp_dataset. There are 4721 connections in which 3798 are "normal" and the remaining 928 connections are "attacks". One of the assumption of our study is, the anomalous connections must be very less compared to the normal connections, without making any modifications it was satisfied in our smtp_data, we reduced the number of attacks in the ftp_dataset to 191, these 191 connections are randomly chosen, and the new dataset is called as ftpnew.

We implemented the Hierarchical clustering algorithm using all the selected distance measures and applied SNC for labeling. With ftpnew dataset the Hierarchical clustering (or PAM) with Euclidean distance, Manhattan distance, Minkowski for p values 0.75 and 1.5, the results were so poor that the detection rate is almost zero. The 'attack' connections were in big clusters which are labeled as normal by SNC. Hierarchical clustering with Bray/Curtis, Ruzicka and Roberts distance has produced better results. The best result is detection rate of 84.2% at a false alarm rate of 4.6%, this is achieved by hierarchical clustering with Bray/Curtis. The confusion matrix of this is given in table IV. PAM also performed well with the Bray/Curtis, Ruzicka, and Roberts distance measures, the best among them was detection rate of 100% with false alarm rate of 12.7% achieved with Bray/Curtis distance. While using PAM with Bray/Curtis and k=10, SNC labeling algorithm was able to label all the 191 attacks correctly, but mislabeled 483 normal instances as attacks.

TABLE IV. CONFUSION MATRIX AFTER APPLYING SNC LABELING ALGORITHM

	Normal	*Attack*	
Predicted Normal	3620	24	3640
Predicted Attack	178	167	339
	3798	191	

[d.] ftpdataset using hclust and bray/Curtis distance method

By these experiments we observed that when we partition the dataset in to two clusters expecting the clusters to be Anomaly and Normal, even all the Anomaly connections fall into one partition along with Anomalies many normal connections are also in that cluster which if we treat as Anomaly cluster almost 30 to 40 percent of normal connections will be labeled as Anomaly which is ineffective. As we increased the number of clusters the size of the clusters decreased and Anomalies almost always were in the smallest cluster and SNC labeling algorithm identified that smallest cluster as the Anomaly and other clusters as normal. As we increased the number of clusters in the ftpnew dataset, it formed many small clusters which contains normal instances only, and these clusters are far from big normal clusters, to thwart with this case we used another labeling mechanism i.e. "labeling clusters using class representative objects". In this method we pass the class representative objects whose class is clearly known (in our case we passed 5% of anomaly connections), along with the dataset that is to be clustered as the input to the clustering algorithm. When the clusters are formed we trace in which cluster our class representative objects lie, and treat that whole cluster as similar to the class representative object, so label the whole cluster with the label of class representative object (anomaly) and all the clusters which are smaller than that cluster and far from the biggest cluster (as in SNC labeling algorithm) are also labeled as Anomaly. This improved the results in labeling clusters.

TABLE V. DR AND FAR ON FTPNEW USING REPRESENTATIONAL OBJECT LABELING ALGORITHM

Algorithms	*Distance Measures*	*FAR(%)*	*DR(%)*	*#Clusters*
Hierarchical Clustering Algorithm (hclust)	Euclidean Distance	8.5	100	27
	Manhattan Distance	8.8	100	20
	Minkowski Distance (p=0.75)	8.6	100	25
	Minkowski Distance (p=1.5)	8.6	100	25
	Bray/Curtis Distance	4.5	100	20
	Ruzicka Distance	13.3	100	10
	Roberts Distance	13.3	100	10
K – Medoids (PAM)	Euclidean Distance	9.3	100	25
	Manhattan Distance	8.6	100	25
	Minkowski Distance (p=0.75)	8.6	100	25
	Minkowski Distance (p=1.5)	8.9	100	25
	Bray/Curtis Distance	8.6	100	20
	Ruzicka Distance	8.4	99.4	20
	Roberts Distance	7.9	99.4	10

[e.] Ftpnew dataset

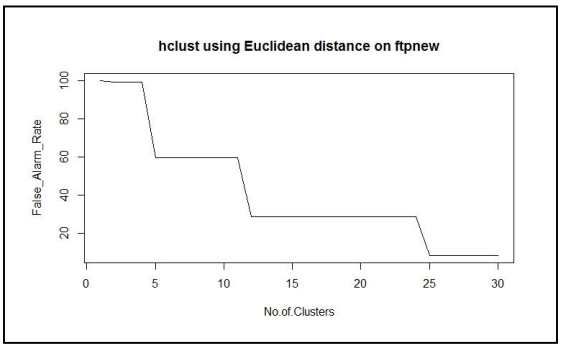

Fig. 1. FAR Vs K

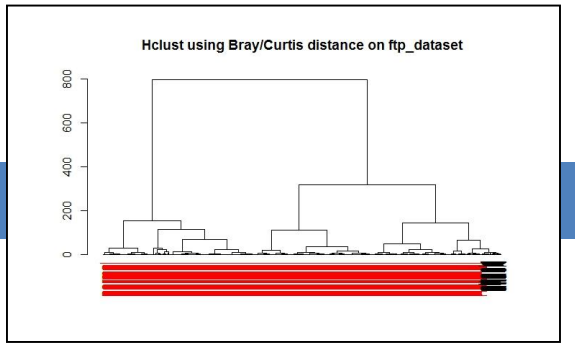

Fig. 2. Hierarchical Clustering on ftpnew using Bray/Curtis

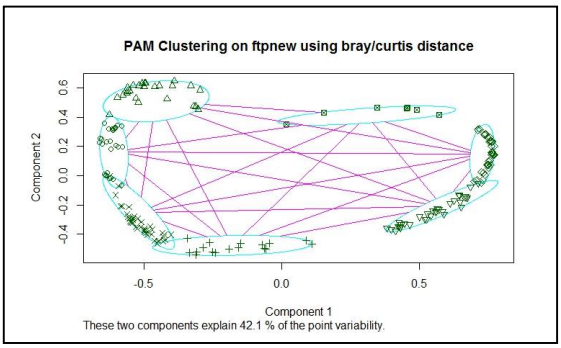

Fig. 3. Clustering on ftpnew using PAM with Bray/Curtis

Fig.1. shows the relationship between the "number of clusters" and FAR, we increased the k (number of clusters) value from 2 to 30 and observed the change in FAR. FAR decreased with the increase of k, up to 30 clusters, then it remained constant until the k is 60, when the k value is further increased, detection rate started to fall. Fig.2. represents the tree like structure which shows hierarchical clusters; this is by using hclust with Bray/Curtis distance measure. In the leaves of the tree all the red ones represent normal and the black ones represent attacks. Fig.3. represents the clusters of ftpnew dataset; the algorithm used is PAM with Bray/Curtis distance measure. The figure shown represents only 200 connections from the ftpnew as it will be clumsy and un-understandable if the whole dataset is shown.

VI. Conclusion

This study emphasizes the utilization of different distance measures in clustering the Network Connections, instead of just sticking to Euclidean distance. The results also showed that Bray/Curtis distance measure performed well compared to other distance measures we considered in the study. Another important consideration of this study is labeling the clusters of network connections.

We used a mechanism for labeling the clusters by using class representative objects which labeled attacks in the KDDCup 99 dataset. This mechanism is like semi supervised learning; along with the objects to be clustered we add some class representative objects and give them as input for clustering algorithm. After the clustering is done we will trace class representative objects and based on these class representative objects the clusters are labeled. This approach worked well in labeling the network connections. This approach cannot find the anomalies which were very similar to the normal connections and this is not practical to apply for real time intrusion detection system, this is a serious drawback.

As the future work we want to study labeling mechanism further and apply them for anomaly detection in various domains i.e. for different datasets. We want to study and find out how different similarity measures effect the formation of clusters.

References

[1] Zulaiha Ali Othman, Azuraliza Abu Bakar, Afaf Muftah Adabashi and Zurina Muda, "A Similarity Normal Clustering Labeling Algorithm for Clustering Network Intrusion Detection," Journal of Applied Sciences, 2014, pp. 969-980.

[2] Leonard Kaufman, and Peter J. Rousseeuw, "Clustering by means of Medoids," Statistical Data Analysis based on L1 Norm and Related Methods, 1987.

[3] M M Deza, and Elena Deza, "Encyclopedia of Distances," Springer-Verlag Berline Heidelberg.

[4] Mahbood Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of KDD CUP 99 Data Set," In Proceeding of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications, (CISDA 2009).

[5] S. J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan, "Cost-based modeling for fraud and intrusion detection: Results from the jam project," disced, vol. 02, p. 1130, 2000.

[6] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup 99.html, October 2007.

[7] E. Eskin, A. Arnold, M. Prerau, L. Rortnoy and S. Stolfo, "A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data,"In Proceedings of Applications of Data Mining in Computer Security. Kulwer Academics, pp. 78-100.

[8] David J. Weller-Fahy, Brett J. Borghetti, and Angela A. Sodemann, "A Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection", IEEE Communication Surveys & Tutorials, Vol 17, No.1, First Quarter 2015, pp. 70-91.